Special Section
Current and Emerging Methods in the Rhetorical Analysis of Texts

Opening: Toward an Integrated Approach

Cheryl Geisler

Simon Fraser University | Canada

Abstract: The rise of digital humanities has led many writing researchers to consider using digital tools to analyze rhetorical patterns in text. Yet taking a digital approach to the analysis of texts is a complex task. We are faced with a variety of techniques and tools, all of which require significant investment to learn and use. How can we best understand the costs and benefits of adopting a particular approach? Are they simply alternatives or can they be integrated? The three sets of authors in this special section attempt to address these questions by using alternative methodologies to analyze a common set of documents. The following opening piece serves as an introduction to the project. In it, we place their research in the context of taxonomy of approaches to text analysis, and review prior attempts at integration. Following the articles, a closing piece examines the prospects for integration. In it, we provide a brief review of the results of the analyses followed by an examination of their commonalities and variations. Finally, we conclude with the considerations that should be taken into account in choosing a method for textual analysis, as well as a discussion of the potential for an integration of methods.

Keywords: text analysis, text mining, rhetorical analysis, data coding, research methodology

1. Introduction to the special section: Current and emerging methods in the rhetorical analysis of texts

The rise of digital humanities has led many writing researchers to consider using digital tools to analyze rhetorical patterns in text. The term digital humanities covers much more than the analysis of texts, of course (Berry, 2012). In the introduction to their recent collection, Ridolfo & Hart-Davidson (2015) have argued for using a digital toolbox for all sorts of rhetorical purposes. And the recent special issue curated by Enoch and Gold (2013) focuses on the creation rather than the analysis of digital archives. Nevertheless, from the beginning, one of the core concerns for those in the digital humanities has been the analysis of texts (Bradley, 2004; Burrowes, 2004; Ide, 2004).

Even a cursory review of the approaches included in these recent pieces suggests that taking a digital approach to the analysis of texts is a complex task. We are faced with a variety of techniques and tools, all of which require significant investment to learn and use. How can we best understand the costs and benefits of adopting a particular approach? Are they simply alternatives or can they be integrated? The authors in this special section on Current and Emerging Methods in the Rhetorical Analysis of Texts propose to address these questions by applying a variety of analytic techniques to a common set of documents.

The documents examined were produced by eight pairs of PhD advisors and their advisees across four disciplines (Computer Science, Chemical Engineering, Materials Science Engineering, and Humanities and Social Sciences) as part of a larger study of source usage conducted by Karatsolis (2005, 2011). In that earlier work, the documents were used as the basis for discourse-based interviews in which authors were shown specific instances of source use from their own documents and then were asked whether using another citation pattern would have made a difference. Authors were also asked to describe the typical way they incorporated sources into their reading and writing processes. The main purpose of that study was to understand the decision-making processes that academics used to incorporate sources into their work, and the primary analysis concerned itself with the interview data (Karatslis, 2011).

For the purposes of the work reported in this special section, the authors returned to more closely examine the documents gathered in that study. Each research group received half of the documents with Karatsolis' original coding and half without any coding. Karatsolis' coding schemes were also shared. More details about the data and the original coding are reported by Karatsolis in the first of the articles in this special section. Since we were interested in better understanding what could be learned using each of the methods, the analyses undertaken by the other two groups were not limited to Karatsolis' original research questions, although all understood that source usage was an important focus. All of the research groups received ethics approval from their home institutions for this secondary use of data.

In the articles that follow, readers will find a presentation of results. Karatsolis presents the results of the original hand-coded analysis of citation patterns. Kaufer, Ishizaki, and Chi review what they discovered using DocuScope, a text analysis environment with a suite of interactive visualization tools built for rhetorical analysis. And Omizo & Hart-Davidson report on the results of using machine-learning techniques to locate and classify signals that mark common rhetorical moves. Following these three analytic reports, I return with an exploration of the analytic variations that lie behind their results. With this synthesis in hand, we will then be in a better position to address the questions with which we begin: How can we best understand the costs and benefits of adopting a particular approach? Are they simply alternatives or can they be integrated?

2. Taxonomy of Analytic Methods

Before we can understand the range of analytic techniques that our authors have employed, we need a more general overview of approaches to text analysis. The taxonomy presented here took as its starting point work by Pollach (2012), intended to introduce to the techniques of corpus linguistics to the field of computer-aided text analysis. In her work, Pollach distinguishes between three broad classes of approaches as I have diagrammed in the middle of Figure 1: computer-aided content analysis, computer-aided interpretive textual analysis, and corpus linguistics. To Pollach's taxonomy, I have added, on the far right, text mining, an approach arising out of the data mining tradition, and, on the far left, hand coding which uses human coders rather than digital processing.

Briefly and starting from the left in Figure 1, hand-coding (shown in very light grey) refers to the analytic process through which humans assign textual segments to coding categories guided by the definitions supplied in a coding scheme. Next to it, computer-aided content analysis (shown in light grey), refers to the process of automatically coding textual segments using content dictionaries that map words and phrases onto content categories. In the lower middle of Figure 1 (shown in medium grey), computer-aided interpretive textual analysis, also known as computer-assisted qualitative data analysis or CAQDAS, takes an interpretive stance, assisting but not replacing human coders, by helping them to manage, retrieve, code and link data. Corpus linguistics (shown in dark grey) examines real world corpora looking for words or terms that co-occur more often that would be expected by chance. Text mining (shown in black on the far right) uses Natural Language Processing (NLP) techniques to create clusters of text attributes in order to distinguish between those texts that have and those that do not have some characteristic recognized by human readers.
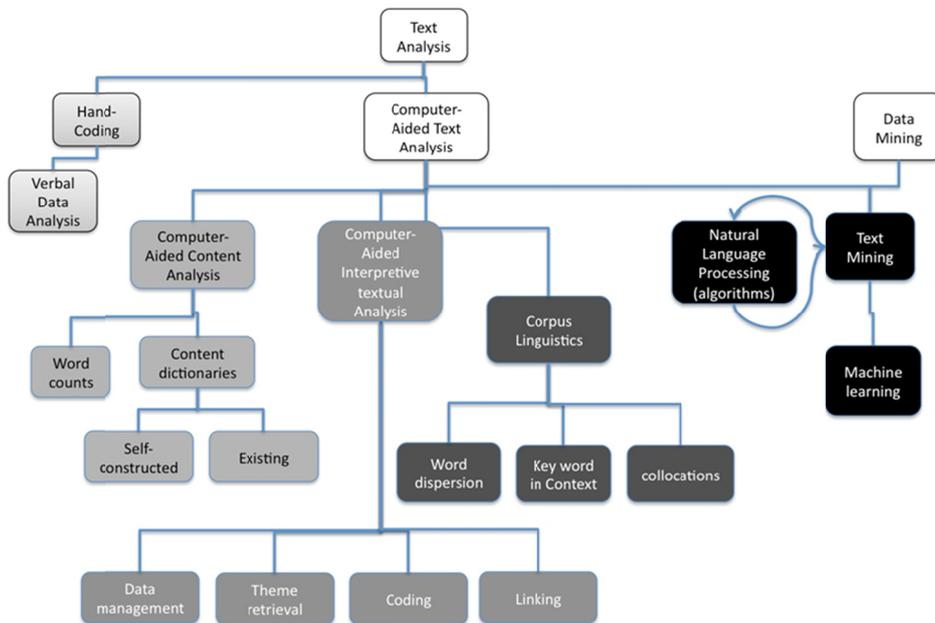
Figure 1: Taxonomy of approaches to text analysis based in part on Pollach (2012).

A review of recent work using digital tools to analyze rhetorical patterns finds the entire scope of this taxonomy in use in writing research. The hand-coding approach was systematized as verbal data analysis in Geisler (2004) and has been used in a variety of research studies (Sun, 2006; Amidon & Blythe, 2008; Swarts, 2011). Hand coding is also often used as part of text mining approaches when human readers are asked to hand code the data that is then used to train the learning algorithms. Graham et al. (2015) also reported using hand coding as the first phase of statistical genre analysis. In this special section, Karatsolis used hand coding in the original analysis of the common data set, and Omizo and Hart-Davidson rely on his hand coding as part of their machine learning approach.

Within the tradition of computer-aided content analysis, Hart and colleagues (2001, 2013, 2015) have been exploring the use of dictionary-based analysis of large corpora. In this special section, Kaufer, Ishizaki, & Chi use a dictionary-based approach as well. Computer-aided interpretive textual analysis has traditionally been associated with qualitative approaches at odds with the more statistical methods used by the other computer-assisted approaches in our taxonomy. An increasing number of software packages, however, are including both qualitative and quantitative tools in a mixed methods package. The use of CAQDAS packages such as Nvivo, QDA Miner, and

Dedoose is very common in writing research. Though not represented in this special section, mixed methods CAQDAS packages may foreshadow the possibility of an integrated approach, a topic to which we'll return in the next section.

Techniques in corpus linguistics have informed a variety of studies in writing. In 2010, Schlitz (2010) edited a special issue of *The Journal of Writing Research* on a corpus-informed approach to writing research. Papers included Hüttner's (2010) comparison of a small-scale purpose-built corpus of student conclusions with a reference corpus; Bloch's (2010) comparison of students' use of reporting verbs to that of published writers; and Henderson & Barr's (2010) preliminary work examining students' use of authorial stance compared to published texts. In later work, Dixon and Moxley (2013) used a concordance tool (AntiConc) to examine teacher's comments on student papers, looking for evidence that they were using a program-wide rubric. Aull (2015) recently used corpus linguistic methods to study the relationship between the language of writing prompts and the genre choices of first-year writers. Although the techniques of corpus linguistics were not brought to bear in the analyses in this special section, Swales' (2014) recent corpus-based study of the citation practices of undergraduate and graduate students in biology is a nice complement to the work reported in this special section.

Text mining may be the most recent addition to the arsenal of tools for the analysis of texts in writing research. In the 2010 volume of this journal, Xiong and colleagues (Xiong, Litman & Schunn, 2012) explored the use of natural language processing techniques to automatically detect the presence or absence of helpful elements in peer feedback on writing. In our special section, Omizo and Hart-Davidson take a text mining approach.

3. Possibility of an Integrated Approach

As I mentioned earlier, the possibility of combining different approaches to text analysis into a common toolkit was one of the motivators of our project. Are the approaches laid out in Figure 1 simply alternatives or can they be integrated? By integration, we mean to draw attention to the possibility that techniques developed in what are often non-overlapping disciplines might be combined in novel and fruitful ways. The idea of an integrated approach is not a new one and I'll detail a few of the more noteworthy integrations here.

In 2009, Janasik and colleagues from the Helsinki University of Technology made one of the first arguments for the integration of text mining techniques with qualitative approaches. Using self-organizing maps (SOM), they show how qualitative researchers can improve the quality of their inferences when examining large data sets, either to help develop relevant concepts in a data-driven approach, or to test the adequacy of existing concept in a theory-driven approach.

The move toward an integrated approach got a big push a few years later in what has become known as the KWALON Experiment. In 2011, Silver from the CAQDAS

Networking Project at the University of Surrey was among a group that invited developers of computer-aided qualitative data analysis software (CAQDAS) across Europe to participate in an experiment that has some similarities to the project in this special issue. Each group was given a large data set for analysis and invited to present results of the analysis at the KWALON conference. The results were also published in the *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research* (Evers, et al., 2011). The experiment involved the use of ATLAS.ti, Cassandre, MAXqda, NVivo, and Transana, all of which would be classified as computer-aided interpretative text analysis applications in Figure 1.

The dataset was large and the only researcher that managed to analyze it all was Lejeune using Cassandre, a free open source text analysis software tool (available online at http://www.cassandre.ulg.ac.be with directions in French). In his report on the experiment (2011), Lejeune describes Cassandre as combining an automatic process which uses concordance functionality to select texts for analysis and an interpretive process in which "the researcher has to comprehensively read, empathize, reflect, and interpret" the selected texts. It was probably this sequencing of automatic and interpretive processes that made Cassandre the only application able to analyze the entire KWALON corpus. His success suggests an opportunity to use hand coding in a later stage of analysis rather than in the early stages as is more common.

Integration has been given a further boost by researchers coming out of the University of Leipzig. Wiedemann (2013), drawing on Moretti's work (2000), argued for the integration of "distant reading" techniques associated with text mining with "close reading" supported by more qualitative methods. Key to this integration, he argues, is the ability to maintain the link between the qualitative input and the quantitative output. In other words, one must be able to inspect text-mining results and, from those results, dig back into the text data in its full context

The Leipzig Corpus Miner, worked on by Wiedemann, Lemke, and other colleagues (Lemke, et.al. 2015), uses what they call Blended Reading, a three stage process. In stage 1, researchers identify relevant documents to make up a corpus using tools like topic models and co-occurrence analysis "to derive or sharpen categories of interest for the upcoming analysis." In stage 2, researchers manually annotate texts using snippets within the most relevant documents identified in stage 1, using their area expertise and checking interrater reliability. The third stage involves using automatic detection, using the annotations produced in the second step, over multiple iterations to improve the quality of the detection. The resulting automatic classifier can then be used to identify appropriate snippets in the corpus. What is intriguing about the approach taken by Lemke and his colleagues is that it seems to combine high quality hand coding with automatic methods.

In what follows we turn to the reports of the three research groups who, in the spirit of KWALON experiment, tried their hand at analyzing the common dataset originally collected by Karatsolis (2005, 2011). Representing a selection of methods surveyed in

Figure 1, they invite us to consider how our choice of analytic methods both invites and constrains our explorations of the rhetorical patterns in texts.

References

Amidon, S., & Blythe, S. (2008). Wrestling With Proteus Tales of Communication Managers in a Changing Economy. *Journal of Business and Technical Communication, 22*, 5-37. http://dx.doi.org/10.1177/1050651907307698

Aull, L. (2015). Linguistic attention in rhetorical genre studies and first year writing. *Composition Forum, 31*.

Berry, D. M. (2012). *Understanding Digital Humanities.* Retrieved from http://dx.doi.org/10.1057/9780230371934.0001

Bloch, J. (2010). A concordance-based study of the use of reporting verbs as rhetorical devices in academic papers. *Journal of Writing Research, 2*(2), 219-244. http://dx.doi.org/10.17239/jowr-2010.02.02.7

Bradley, J. (2004). Text Tools. In S. Schreibman, R. Siemens, and J. Unsworth (Eds.), *A Companion to Digital Humanities* (Chapter 33). Oxford: Blackwell

Burrows, J. (2004). Textual Analysis. In S. Schreibman, R. Siemens, and J. Unsworth (Eds.), *A Companion to Digital Humanities* (Chapter 23). Oxford: Blackwell.

Dixon, Z., & Moxley, J. (2013). Everything is illuminated: What big data can tell us about teacher commentary. *Assessing Writing, 18*(4), 241–256. http://dx.doi.org/10.1016/j.asw.2013.08.002

Enock, J., & Gold, D. (2013). Introduction: Seizing the methodological moment: The Digital humanities and historiography in rhetoric and composition. *College English, 76*(2), 105-114.

Evers, J. C., Silver, C., Mruck, K., & Peeters, B. (2011). Introduction to the KWALON Experiment: Discussions on qualitative data analysis software by developers and users. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 12*(1), Art. 40, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101405

Geisler, C. (2004). *Analyzing Steams of Language: Twelve Steps to the Systematic Coding of Text, Talk and Other Verbal Data.* New York: Pearson/Longman.

Graham, S.S., Kim, S-Y., Hartke, D.M. & Keith, W. (2015). Statistical genre analysis: Toward big data methodologies in technical communication. *Technical Communication Quarterly. 24*: 70-104. http://dx.doi.org/10.1080/10572252.2015.975955

Hart, R. P. (2001). 'Redeveloping DICTION: Theoretical Considerations. In M. West (ed.), *Theory, Method, and Practice of Computer Content Analysis* (43-60). New York Ablex.

Hart, R. P. (2015). Genre and automated text analysis: A demonstration. In J. Ridolfo & W. Hart-Davidson (Eds.), *Rhetoric and the Digital Humanities* (152-168). Chicago: University of Chicago Press.

Hart, R. P., Childers, J. P., & Lind, C. J. (2013). *Political Tone: How Leaders Talk and Why.* Chicago: University of Chicago Press. http://dx.doi.org/10.7208/chicago/9780226023298.001.0001

Henderson, A. & Barr, R. (2010). Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research, 2*(2), 245-264. http://dx.doi.org/10.17239/jowr-2010.02.02.8

Hüttner, J. (2010). The potential of purpose-built corpora in the analysis of student academic writing in English. *Journal of Writing Research, 2*(2), 197-218. http://dx.doi.org/10.17239/jowr-2010.02.02.6

Janasik, N., Honkela, T., and Bruun, H. (2009). Text mining in qualitative research: Application of an unsupervised learning method. *Organizational Research Methods, 12*, 436-460.

Karatsolis, A. (2005). *Synthesizing from sources: Patterns of citation use in the academia and implications for the design of electronic reading and writing systems* (Doctoral dissertation). Retrieved from ProQuest, UMI Dissertations Publishing (document 3173267).

Karatsolis, A. (2011). KAIRON: Toward a technology-based pedagogy of source use practice. *The University of the Fraser Valley Research Review*, *4*, 89-99.

Lejeune, C. (2011). From normal business to financial crisis ... and back again. An illustration of the benefits of Cassandre for qualitative analysis. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *12*(1), Art. 24, http://nbn-resolving.de/urn:nbn:de:0114-fqs1101247

Lemke, M., Niekler, A., Schaal, G. S., and Wiedemann, G. (2014). Content analysis between quality and quantity: Fulfilling blended-reading requirements for the social sciences with a scalable text mining infrastructure. *Datenbank-Spektrum, 15*, 7-14.

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *1*, 54-68.

Pollach, I. (2012) Taming textual data: The contribution of corpus linguistics to computer-aided text analysis. *Organizational Research Methods*, *15*(2), 263-287.

Ridolfo, J., & Hart-Davidson, W. (2015). Introduction. *Rhetoric and the Digital Humanities* (1-12). Chicago: University of Chicago Press.

Schlitz, A. A. (2010). Exploring corpus-informed approaches to writing research. *Journal of Writing Research*, *2*(2), 91-98. http://dx.doi.org/10.17239/jowr-2010.02.02.1

Sun, H. (2006). The triumph of users: Achieving cultural usability goals with user localization. *Technical Communication Quarterly*, *15*, 457-481. http://dx.doi.org/10.1207/s15427625tcq1504_3

Swales, J. M.(2014). Variation in citational practice in a corpus of student biology papers: From parenthetical plonking to intertextual storytelling, *Written Communication*, *31*, 118-141, http://dx.doi.org/10.1177/0741088313515166

Swarts, J. (2011). Technological literacy as network building. *Technical Communication Quarterly*, *20*, 264-302. http://dx.doi.org/10.1080/10572252.2011.578239

Weidemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *14*(2), http://nbn-resolving.de/urn:nbn:de:0114-fqs1101247

Xiong, W., Litman, D. & Schunn, C. (2012). Natural language processing techniques for researching and improving peer feedback. *Journal of Writing Research 4*(2), 155-176. http://dx.doi.org/10.17239/jowr-2012.04.02.3

Articles in special section

Karatsolis, A. (2016, accepted for publication). Rhetorical Patterns in Citations across Disciplines and Levels of Participation. *Journal of Writing Research, 7*(3), p. 425-452. doi: 10.17239/jowr-2016.07.03.06

Kaufer D., Ishizaki S., & Cai X. (2016). Analyzing the language of citation across discipline and experience levels: An automated dictionary approach. *Journal of Writing Research, 7*(3), 453-483. doi: 10.17239/jowr-2016.07.03.07

Omizo R.,& Hart-Davidson W. (2016, accepted for publication). Finding genre signals in academic writing. *Journal of Writing Research, 7*(3), 485-509. doi: 10.17239/jowr-2016.07.03.08